# The Battle Against Bots

## Current Threats and New Directions to Counter Automated Attacks

Elisa Chiapponi
elisa.chiapponi@amadeus.com

SoSySec Seminar

Rennes, France

22th November 2024

amadeus
Global Security Operations

# Who am I

- Security Researcher in the **Global Security Operations** of Amadeus
  - Protection of web domains linked to the travel industry
- Expertise in **Network** and **Application** Security
- Work based on Ph.D. and current research and collaborations

EURECOM
Sophia Antipolis

amadeus
Global Security Operations

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Resilient Computing and Cybersecurity Center

SORBONNE UNIVERSITÉ

UNIVERSITY OF TWENTE.

TELECOM SudParis

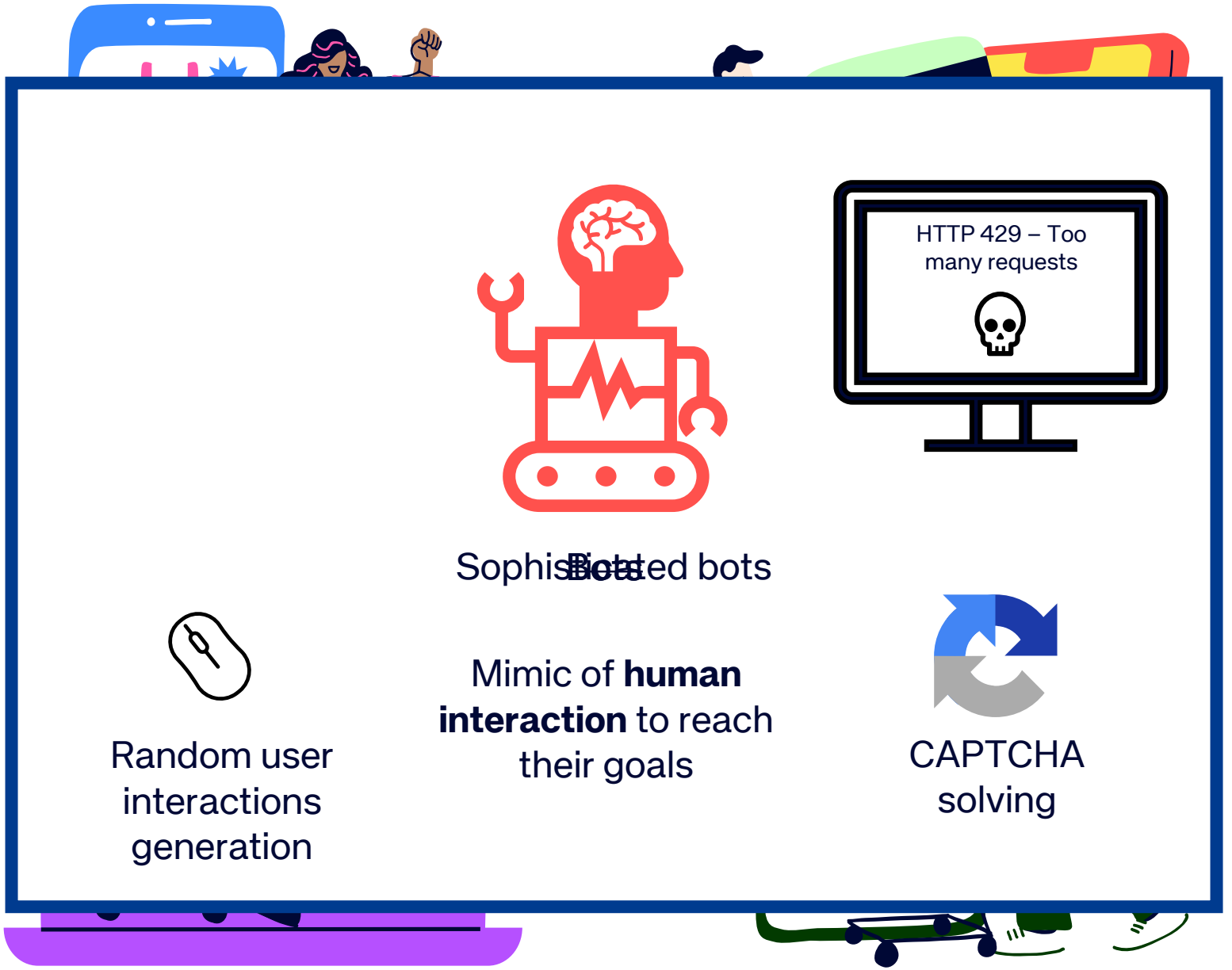
IP PARIS

**RESCUE – Resilient Cloud for Europe**
**IPCEI – Germany**
**Amadeus Germany GmbH**

Supported by:
Federal Ministry for Economic Affairs and Climate Action

Funded by the European Union
NextGenerationEU

on the basis of a decision by the German Bundestag



# Sophisticated Bot Attacks – Functional Abuse


## Web scraping

Competitor Monitoring

Content Reselling

Illicit Aggregators

Requests

Purchases

Revenues

Slow connections


## Denial Of Inventory

Disrupt supply and demand

Holding cheapest fares

Impact revenues and operations

Defeat the competition

Requests

Purchases

Revenues

Reputation


## SMS Pumping

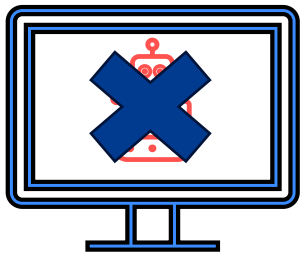
Impact revenues

Generate revenue through network operators

Revenues

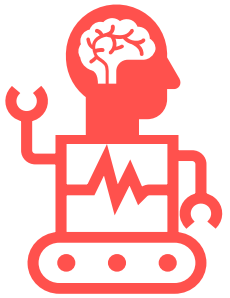
Reputation

# Arms race



New server-side bot detection technique

New bot detection circumvention technique



# What can we do?

Be consta
to learn al
tech

and raise
ess

amadeus
Global Security Operations

# Problems in current landscape

Large usage of Residential Proxies

Redirection of CAPTCHA tests to CAPTCHA Farms

Realistic fingerprints fast rotation

Side effects of current mitigation techniques reveal detection



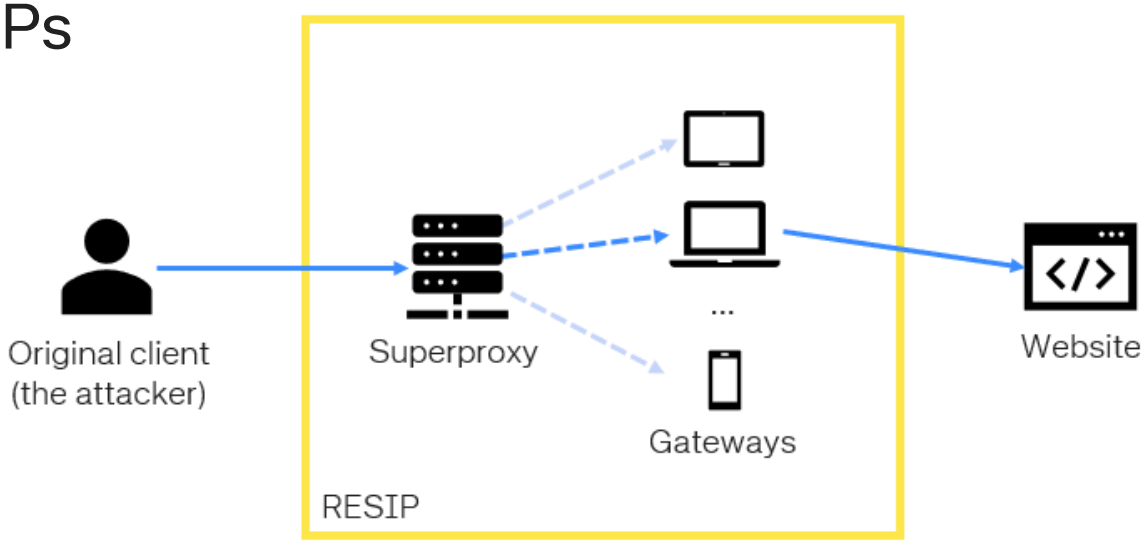


[1] E. Chiapponi (2023). Detecting and Mitigating the New Generation of Scraping Bots. In Ph.D. Dissertation, Sorbonné Université, Cryptography and Security.

# Large usage of Residential Proxies

## The problem

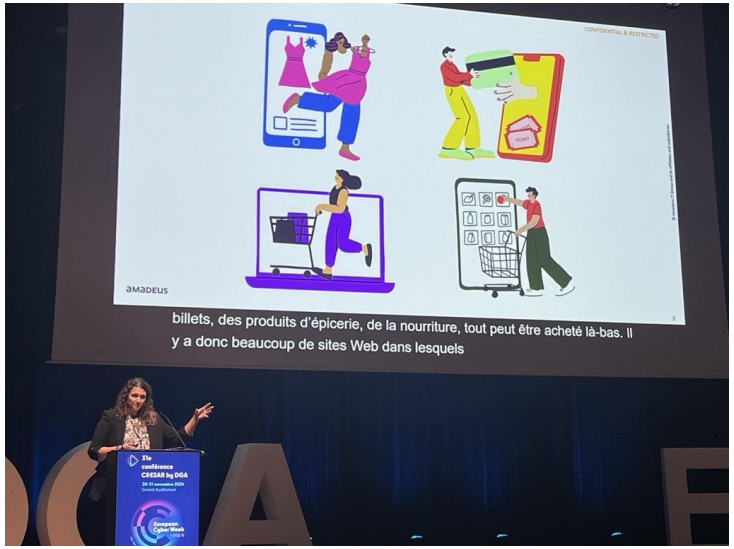
- Large networks of **residential devices** (smartphones, laptops, tablets,...)
- Devices **owned** by genuine users who **share** their usage
- No application layer information about being proxied
  - **Indistinguishable** from the requests sent directly by the residential devices at this layer
  - **High probability of false positives** for the traditional server-side bot detection techniques
- Advanced bot traffic **heavily rely** on RESIPs
- Anyone can build sophisticated bots:
  - Automated **CAPTCHA Solving**
  - Automated **fingerprint rotation**



# Large usage of Residential Proxies

## Our approach

- Study of Residential Proxies (RESIPs) **infrastructure**
  - Identification of **transport layer differences** among direct and proxied connection
  - Leverage of these difference to have two techniques to **detect server-side their usage**
- **Know better** your adversary and **raise awareness** among network operators
  - Testbed to **act as a RESIP gateway**
  - **First study** the **encrypted** traffic the proxied out




[2] E. Chiapponi et al. (2022). BADPASS: Bots taking ADvantage of Proxy AS a Service. In ISPEC 2022.
[3] E. Khan et al. (2024) A First Look at User-Installed Residential Proxies From a Network Operator's Perspective. In CNSM 2024

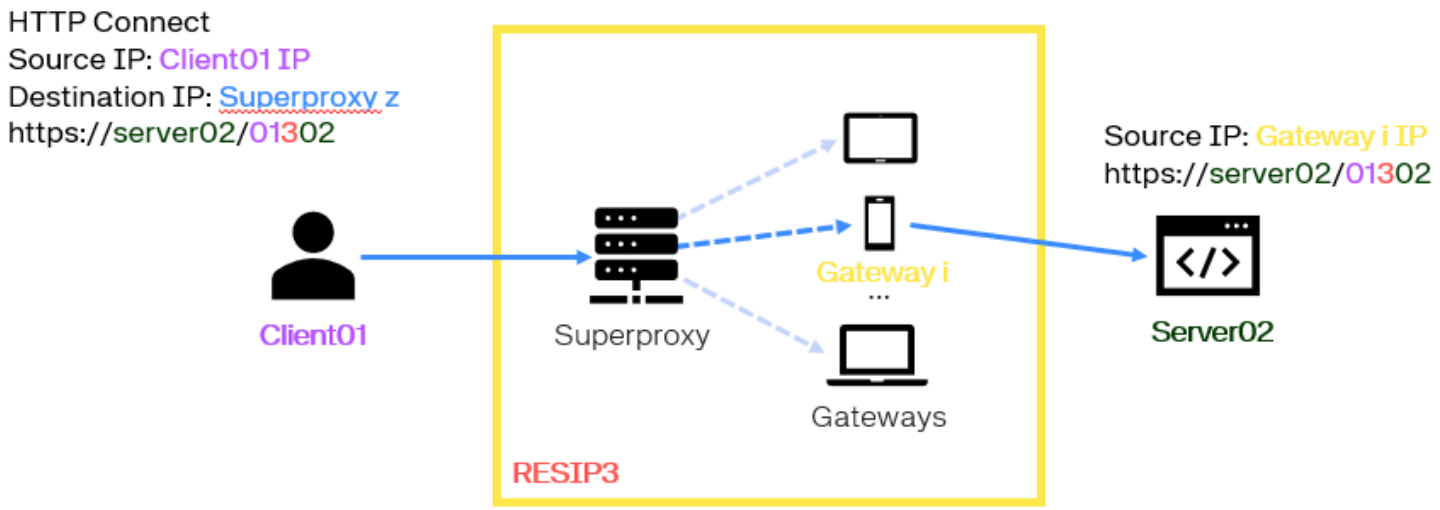
# Large usage of Residential Proxies

## Our approach

- **Arms race** and **limitations** in the techniques
  - We need to **anticipate** the next steps and keep **complementing** our detection
- Keep **studying** the traffic RESIPs proxy out from the testbed
  - Collect **more proofs** of malicious activities
  - Possible find evidence of malicious activity in the **not encrypted traffic**
- Can we know more about the **devices** and the **IPs** in these networks?

# RTT Dataset

- **4** months collection
- **4** RESIP providers
- **2 client/server** machines in **11 locations** in the world
- Requests from each client to each server through **each RESIP network**
- **69M+** RESIP connection








[4] E. Chiapponi et al. (2023). Inside Residential IP Proxies: Lessons Learned from Large Measurement Campaigns. In WTMC 2023.

# Gateways Assignation

1. **Minimization** of gateway IP **repetitions** in a **single client-server path** but not on among all paths

| RESIP | # connections | # countries | # /32 | # /24 | # /16 | # /8 | # ASes | Repeated IPs | Repeated IPs per server | Repeated IPs per client |
|---|---|---|---|---|---|---|---|---|---|---|
| BR | 2,413,405 | 226 | 1,546,886 | 712,274 | 23,274 | 193 | 17,026 | 31% | 3±1.6% | 3.3±1.8% |
| OL | 22,387,788 | 226 | 6,660,452 | 846,165 | 15,230 | 194 | 19,370 | 49% | 16.3%±0.5% | 16.3%±1.3% |
| PR | 22,523,876 | 234 | 3,982,149 | 411,949 | 14,145 | 201 | 9,871 | 61% | 23% | 23.4%±0.2% |
| SM | 22,353,578 | 224 | 6,852,898 | 859,946 | 15,288 | 194 | 19,501 | 49% | 15.7±0.4% | 15.7%±0.4% |

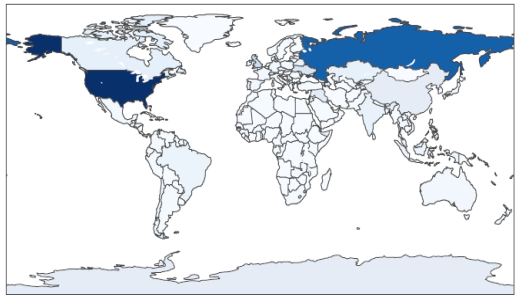




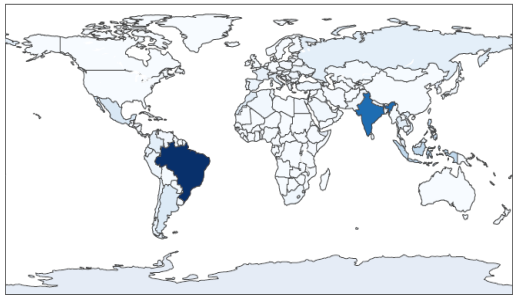
[4] E. Chiapponi et al. (2023). Inside Residential IP Proxies: Lessons Learned from Large Measurement Campaigns. In WTMC 2023.

# Machines distribution

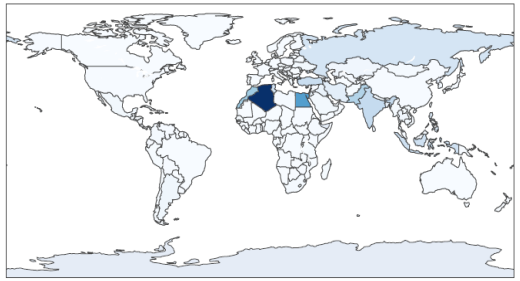
## 2. **Similar gateways geographical** distribution for two providers



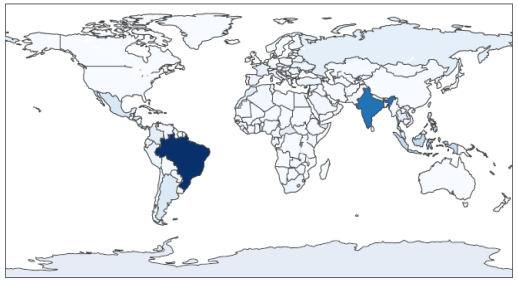
Bright Data



Oxylabs



Proxyrack



Smartproxy

## 3. **Shared IP Pool** among providers

| | BR | OL | PR | SP |
|---|---|---|---|---|
| BR | - | 9% | 5% | 9% |
| OL | 2% | - | 8% | 63% |
| PR | 2% | 13% | - | 13% |
| SP | 2% | 61% | 7% | - |



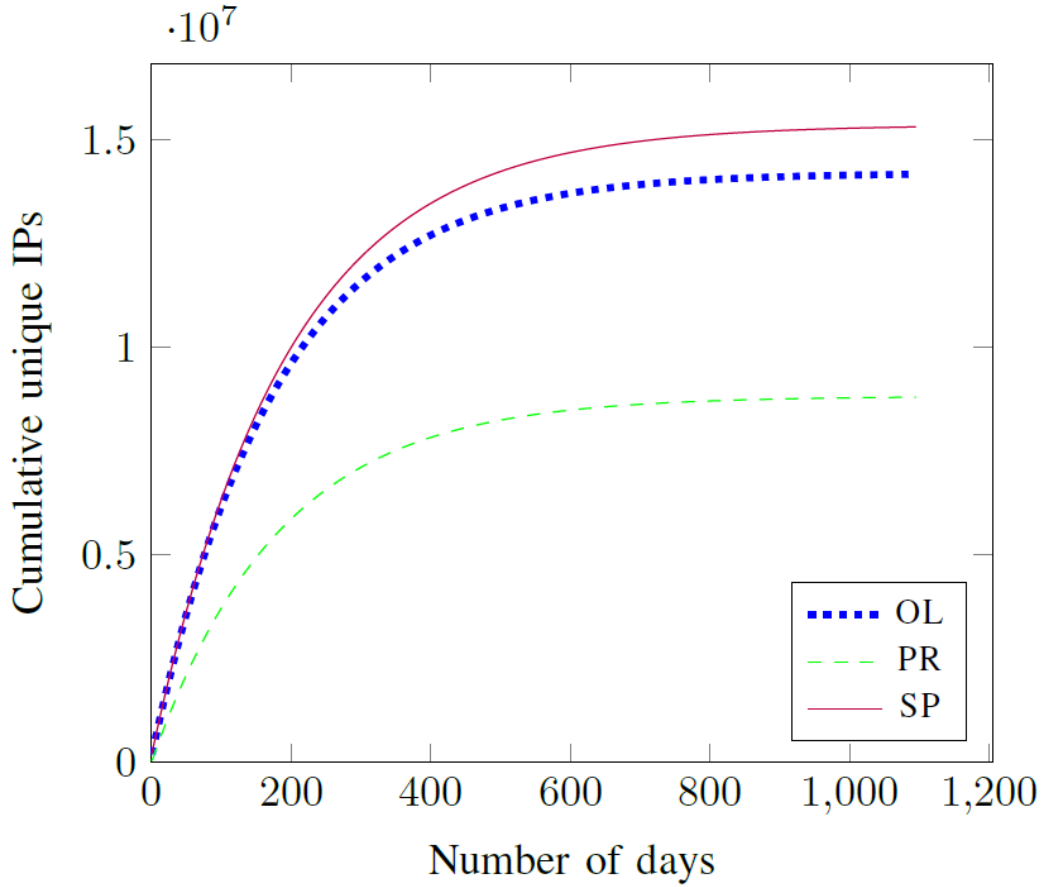
[4] E. Chiapponi et al. (2023). Inside Residential IP Proxies: Lessons Learned from Large Measurement Campaigns. In WTMC 2023.

# Amount of machines

4. Advertised IP pool sizes **do not match** our observation and projections

Assumptions:

- **Constant rate** of devices entering and exiting the network
- **No 1-1 correspondence** between the # of devices and #of IP addresses
  - Generally, **overestimation**
- **Upper bound** for the number of devices








[4] E. Chiapponi et al. (2023). Inside Residential IP Proxies: Lessons Learned from Large Measurement Campaigns. In WTMC 2023.

# External RESIP dataset comparison [1/2]

- External DS 1:
  - May 2017 - March 2018 (vs Jan 2022 – May 2022)
  - 6,419,987 RESIP IPs from 5 RESIP providers
- Sharing two RESIP providers with our study, BrightData and Proxyrack

| DB | IP repetition | IP repetition BD | IP repetition PR |
|---|---|---|---|
| RTT DS | 2.87 % | 2.52 % | 1.26 % |
| External DS 1 | 6.26 % | 0.97 % | 5.86 % |

| DB | /24 repetition | /24 repetition BD | /24 repetition PR |
|---|---|---|---|
| RTT DB | 46.04 % | 33.17 % | 29.15 % |
| External DS 1 | 45.52 % | 19.96 % | 34.74 % |





External DS: https://rpaas.site/

# External RESIP dataset comparison [2/2]

- External DS 2:
  - April 2021 - October 2021 (vs Jan  2022 – May 2022)
  - 9,077,278  Chinese RESIP IPs from 6 RESIP providers

| DB | IP repetition |
|---|---|
| RTT DS | 5.22 % |
| External DS 2 | 8.04 % |

| DB | /24 repetition |
|---|---|
| RTT DB | 54.33 % |
| External DS 2 | 58.52 % |

# What did we learn about the IPs

- Each provider **reuses** IPs among different paths (and possibly users)
- Different providers **share** pools of IPs
- The total amount of RESIP IPs is **smaller** than advertise values
- IP changes, **/24 vary less**
- Can we **leverage** this information?
  - **Tracking /24** and associate the ones where RESIPs appear to a **risk score**
  - Genuine users **share** their devices -> **Whitelisting** to reduce FPs
    - Association of IPs completing a **confirmed human action** (e.g. booking) to the corresponding **fingerprint**
- **Next step:** track the **coverage** with the RESIP IPs detected in Amadeus + complement with study of **number of devices** (Böck, L. et all. (2023). How to Count Bots in Longitudinal Datasets of IP Addresses. 10.14722/ndss.2023.24002.)

# Problems in current landscape

Large usage of Residential Proxies

Redirection of CAPTCHA tests to CAPTCHA Farms

Realistic fingerprints fast rotation

Side effects of current mitigation techniques reveal detection

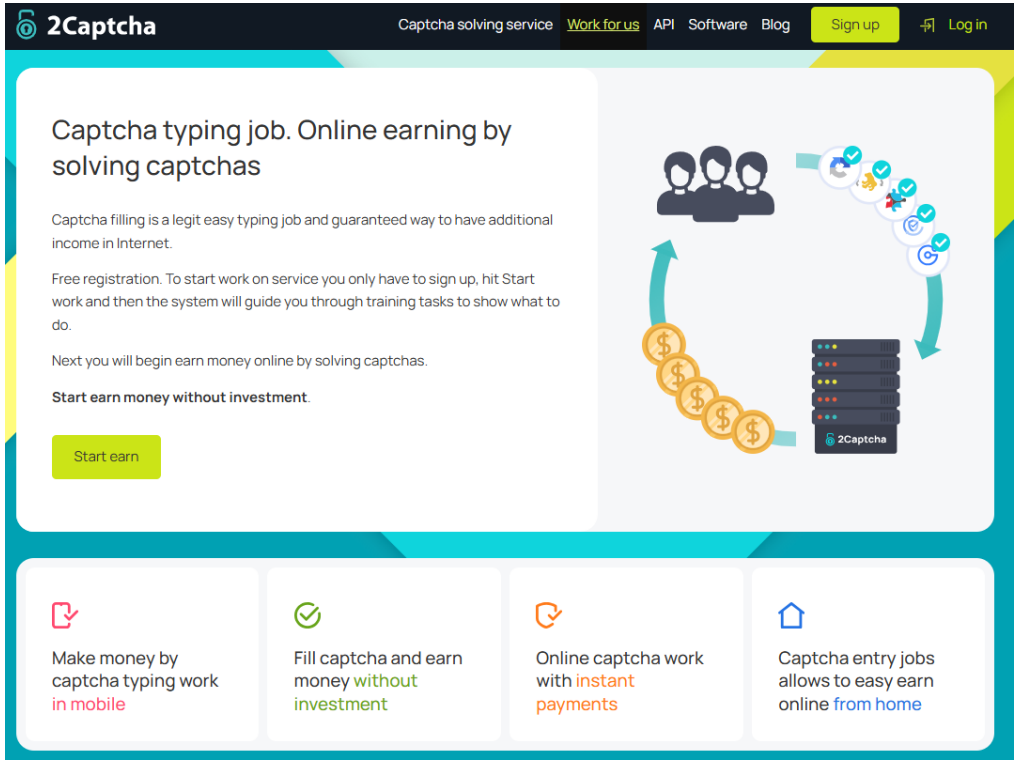


[1] E. Chiapponi (2023). Detecting and Mitigating the New Generation of Scraping Bots. In Ph.D. Dissertation, Sorbonné Université, Cryptography and Security.

# Redirection of CAPTCHA tests to CAPTCHA Farms

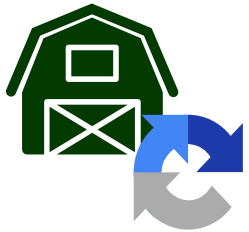
## The problem

- CAPTCHA Farms are **virtual or physical** realities where people are paid to solve CAPTCHA tests **redirected to them**



# Redirection of CAPTCHA tests to CAPTCHA Farms

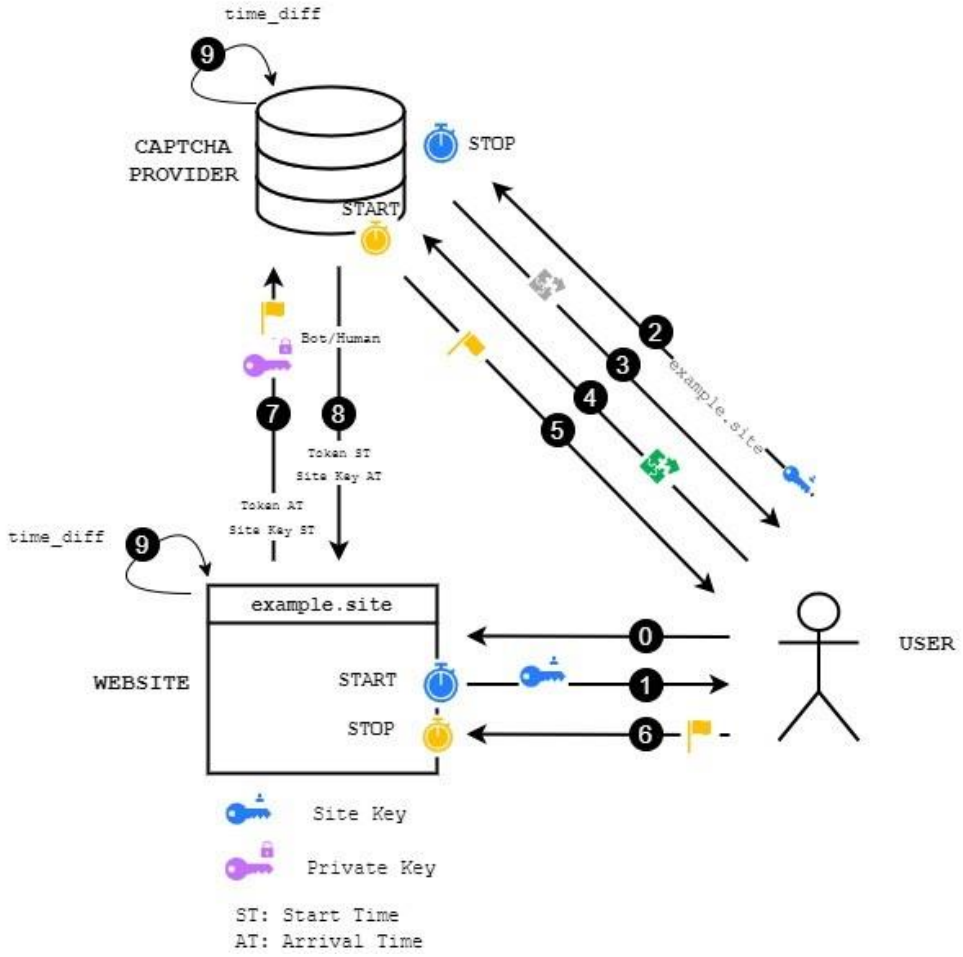
## The problem

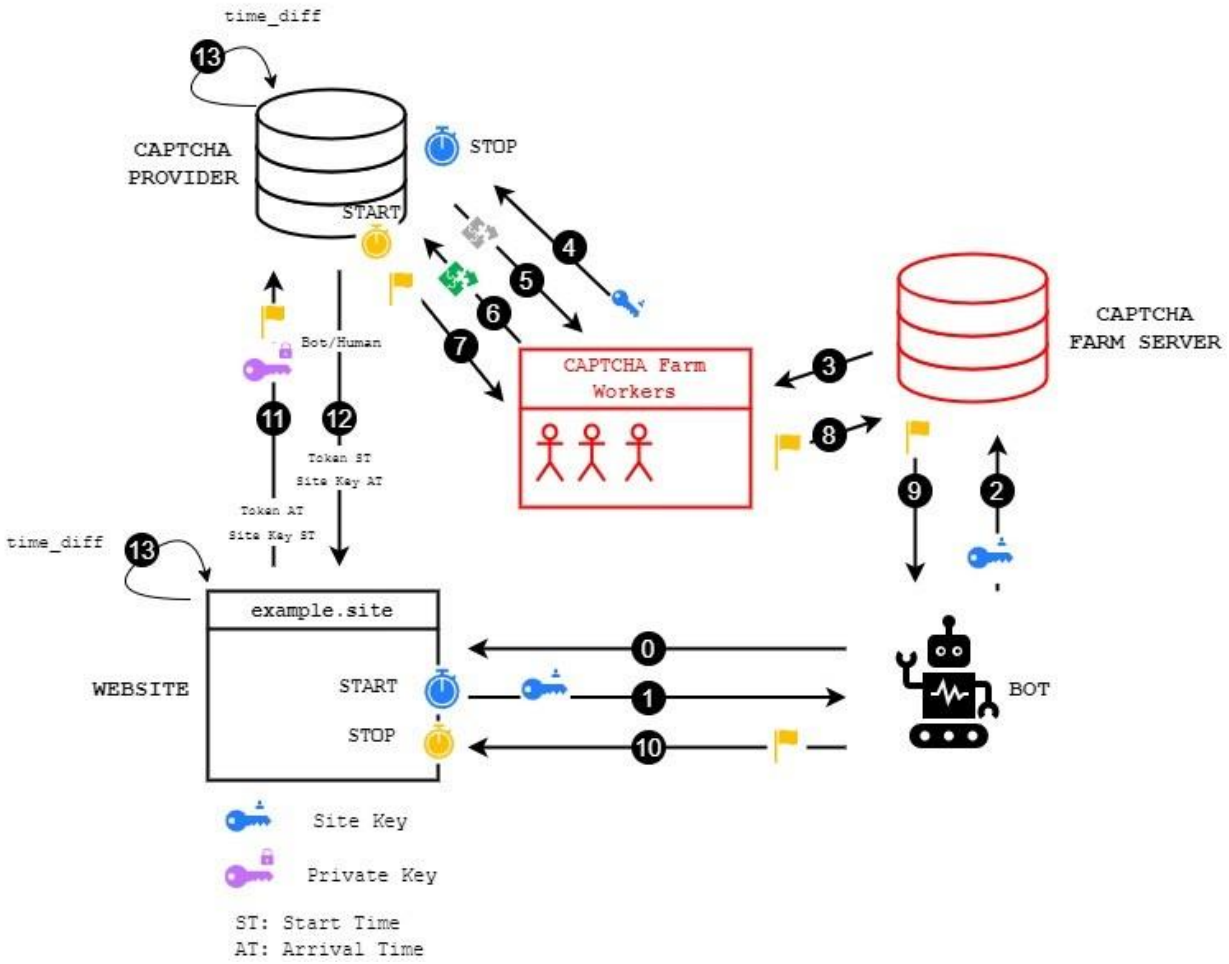


- Due to these farms, CAPTCHA Tests are not an **effective** way to differentiate between human and bots
- CAPTCHA Farms also take advantage of proxies to show the **same IP** and **fingerprint** of the client
- How could we make CAPTCHA Test a **strong mitigation** again?

# Redirection of CAPTCHA tests to CAPTCHA Farms

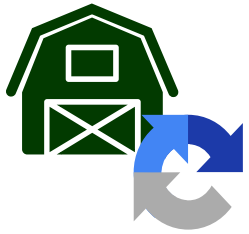
## Our approach



# Redirection of CAPTCHA tests to CAPTCHA Farms

## Our approach



- Site key and token **propagation times** can give an **estimation** of the distance among parties
- We can check if these propagation times are **compatible** with **geographic location** as expressed by the IP of the parties and/or other parameters
  - **Network congestion** need to be taken into account
  - **Possible false positive** for VPNs, VPN IPs whitelisting
- **Stage**
  - POC **Design**

# Problems in current landscape

Large usage of Residential Proxies

Redirection of CAPTCHA tests to CAPTCHA Farms

Realistic fingerprints fast rotation

Side effects of current mitigation techniques reveal detection



[1] E. Chiapponi (2023). Detecting and Mitigating the New Generation of Scraping Bots. In Ph.D. Dissertation, Sorbonné Université, Cryptography and Security.

# Realistic fingerprints fast rotation

## The problem

- The majority of commercial anti-bot solutions are **fingerprint based**
  - **Clustering** bot requests on the same fingerprint/ML model result based on fingerprint and signals
- Nowadays:
  - Bot fingerprints are **difficult to distinguish** from the ones of common real users
  - Sophisticated bots **keep rotating** the fingerprints even when there are not detected already
  - Multiple version of the bot run in **parallel**, one with high volume of traffic, the other ones with low volumes.
    - When the high volume traffic is detected and mitigated, the corresponding version of the bot **stops** its activity and **another version** of the bot **increases the volume** of its traffic
- Detection engine **running constantly** + **analysts' exhaustion**

# Realistic fingerprints fast rotation

## Our approach

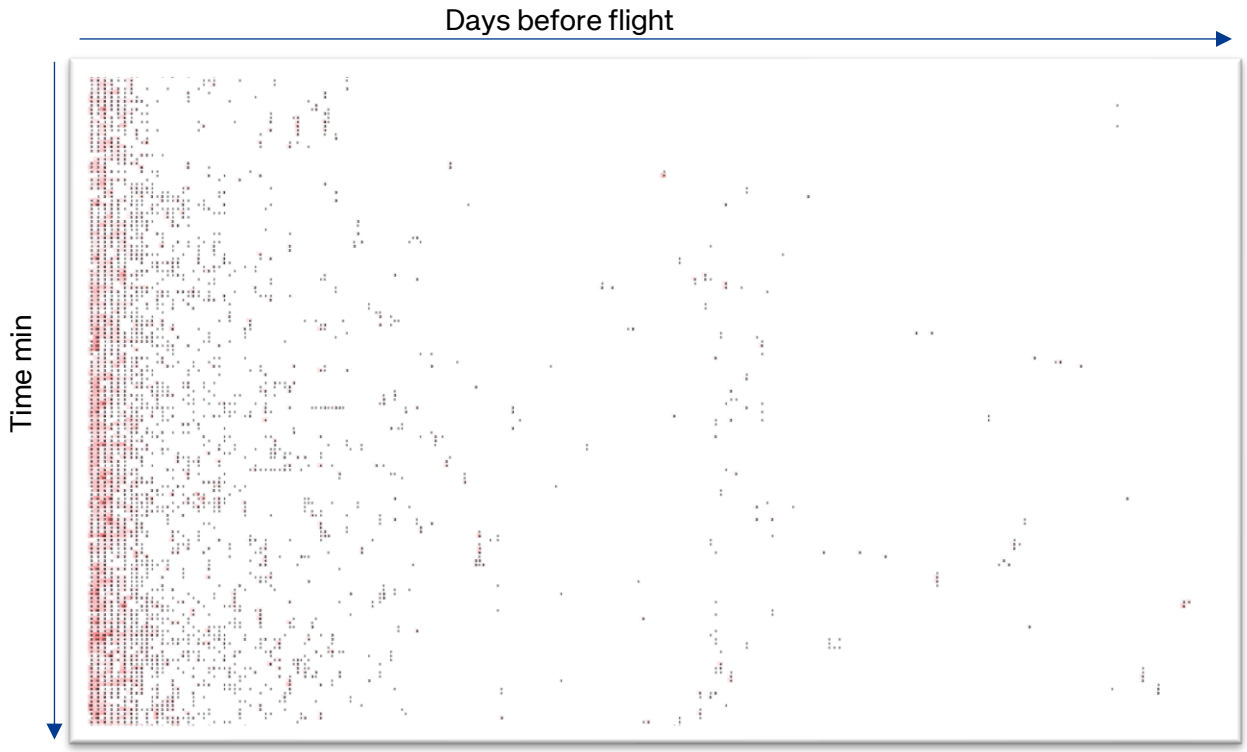
- Bots keep rotating **how technically** they send requests to a website but they will not change **why** they sent those requests
  - Detection based on the interaction on the website and/or requested information

- **Two** approaches under study
  - Bot isolation based on **search patterns**
  - **Graph analysis** of the user interactions

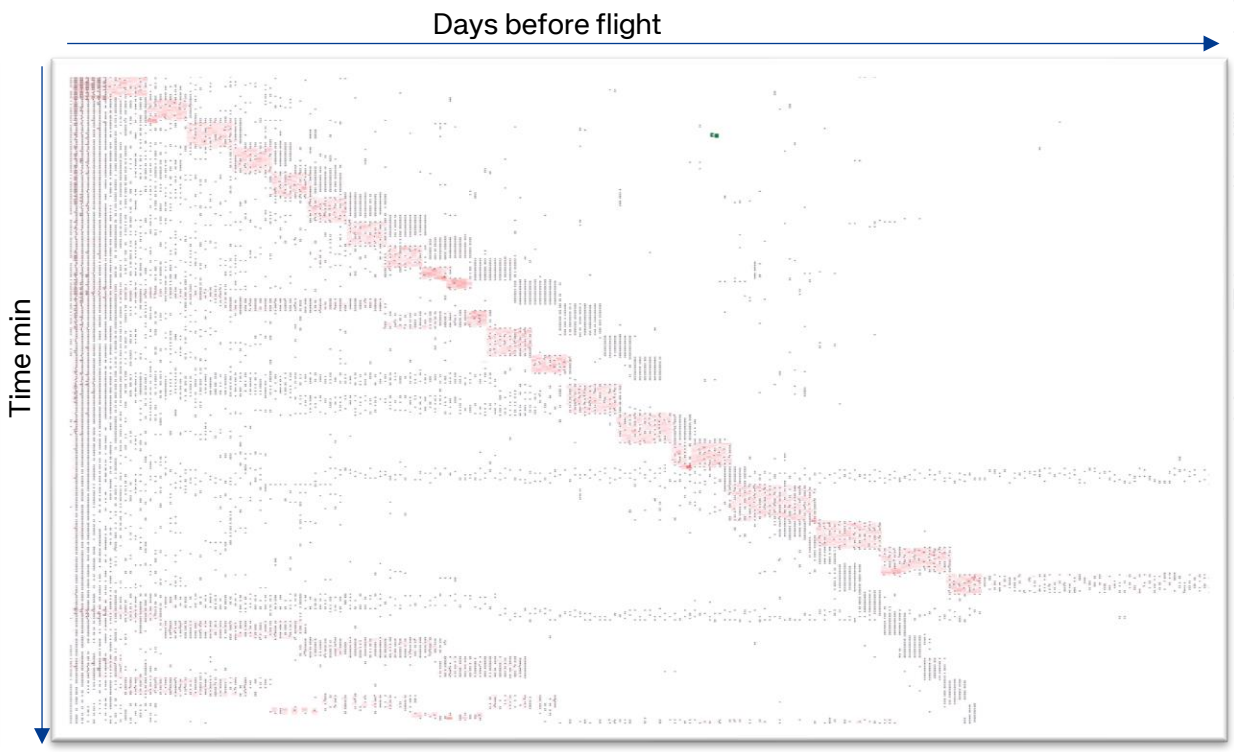
# Realistic fingerprints fast rotation

## Our approach [1/2]

- Bot isolation (also) based on **payload content** that highlight patterns
  - e.g. **Combination** of departure-arrival location and time between the departure date and the date of the search




Normal Traffic




Traffic with bots

# Realistic fingerprints fast rotation

## Our approach [1/2]

- **Advantages**
  - Bots **do not rotate** the parameters of the search
  - **Complementary** to fingerprinting
- **Challenges**
  - Applicable **only** to attacks where there is a search
  - Clearly differentiate real customers from bots to **avoid false positives**
- **Stage**
  - Studying application logs to highlight all possible patterns and **understand the feasibility** of the solution

# Realistic fingerprints fast rotation

## Our approach [2/2]

- **Graph analysis** of the user interaction with the website
  - Graph of all the **possible interaction** on the website
  - **Graph of each user**
  - **Clustering** based on the activity
- **Advantages**
  - Leverage **domain specific knowledge**
  - Detect **attacks** that do not follow under the main ones already considered
- **Stage**
  - **Feasibility** study and initial **testing**

# Problems in current landscape

Large usage of Residential Proxies

Redirection of CAPTCHA tests to CAPTCHA Farms

Realistic fingerprints fast rotation

Side effects of current mitigation techniques reveal detection





[1] E. Chiapponi (2023). Detecting and Mitigating the New Generation of Scraping Bots. In Ph.D. Dissertation, Sorbonné Université, Cryptography and Security.

# Side effects of current mitigation techniques reveal detection

## The problem

- The actors behind the bots can **infer their detection** from the **side effects** of mitigation techniques (blocking, delaying answers,...)
- Once understood they have detected, they **change** approach/fingerprint
- What if we could **prevent them** from knowing they have been detected providing **incorrect but plausible answers**?



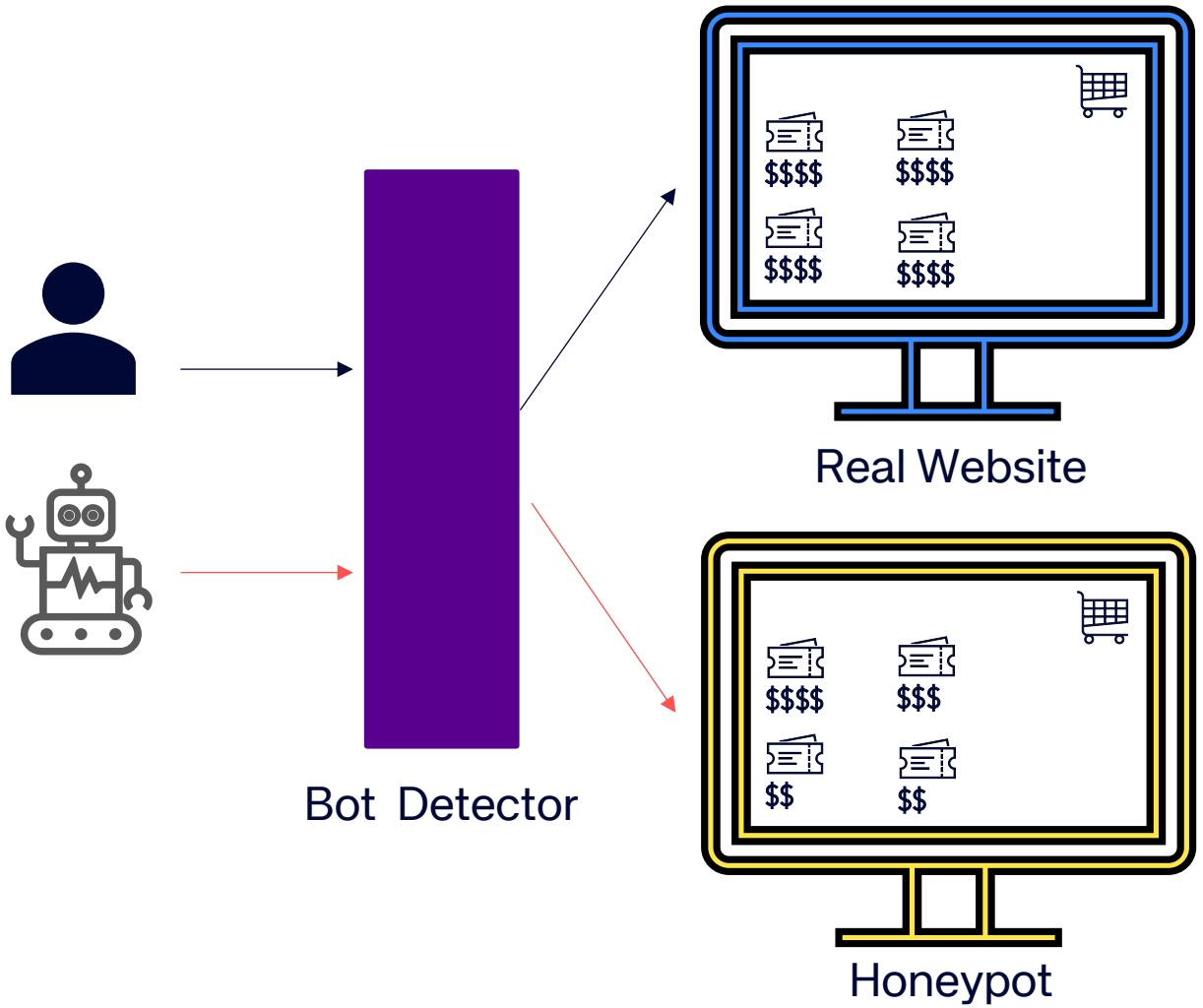


[5] E. Chiapponi (2021). Scraping Airlines Bots: Insights Obtained Studying Honeypot Data. In International Journal of Cyber Forensics and Advanced Threat Investigations.

# Side effects of current mitigation techniques reveal detection

## Our approach

- **Redirection** of detected bots to a honeypot **mimicking the real website**
- **Luring** of the bots that do not have a direct feedback of detection
- Database **poisoning**








[5] E. Chiapponi (2021). Scraping Airlines Bots: Insights Obtained Studying Honeypot Data. In International Journal of Cyber Forensics and Advanced Threat Investigations.

# Honeypot to lure the attackers

- Initial **POC** in 2020
- Collaboration with an airline company, redirection of **specific bot signature**
- Running for **56 days** (interruption linked with COVID-19 restrictions on flights)
- After 3 days from the start of the case study, **modification of fares**: increase the real price by 5% for 10% of the requests
- Amount and timing of the requests **in line** with those before the honeypot
- Bots were **not sophisticated enough** to detect small changes





[5] E. Chiapponi (2021). Scraping Airlines Bots: Insights Obtained Studying Honeypot Data. In International Journal of Cyber Forensics and Advanced Threat Investigations.

# Honeypot to lure the attackers

- **Advantages**
  - **Poisoning** the fare dataset of the attacker
    - **Reduction** of economic incentive in attacks
  - **Increasing** the cost of bot attacks(additional checks to identify honeypot responses)
- **Opportunity**
  - **Expand** the concept to Denial of Inventory attacks
- **Challenges**
  - Fare retrieval and associated **costs**
    - Cache, ML generation, ad-hoc algorithm
- **Stage**
  - Feasibility and **cost assessment**

# Problems in current landscape

Large usage of Residential Proxies

Redirection of CAPTCHA tests to CAPTCHA Farms

Realistic fingerprints fast rotation

Side effects of current mitigation techniques reveal detection



[1] E. Chiapponi (2023). Detecting and Mitigating the New Generation of Scraping Bots. In Ph.D. Dissertation, Sorbonné Université, Cryptography and Security.

# How are we addressing the problems in the current landscape?

- Server-side detection based on **transport layer differences**
- **Study** of the traffic proxied out by the gateways
- /24 **reputation** DB

- CAPTCHA Farm redirection based on **propagation time** of elements exchanged by the involved parties

- Bot isolation based on **search patterns**
- **Graph analysis** of the user interactions with the website

- **Honeypot** reproducing the real website

# Thank you for your attention!

**More questions?**
**elisa.chiapponi@amadeus.com**
**or here in person**

Presentation based on:

[1] E. Chiapponi (2023). Detecting and Mitigating the New Generation of Scraping Bots. In Ph.D. Dissertation, Sorbonné Université, Cryptography and Security.
[2] E. Chiapponi et al. (2022). BADPASS: Bots taking ADvantage of Proxy AS a Service. In ISPEC 2022.
[3] E. Khan et al. (2024) A First Look at User-Installed Residential Proxies From a Network Operator's Perspective. In CNSM 2024
[4] E. Chiapponi et al. (2023). Inside Residential IP Proxies: Lessons Learned from Large Measurement Campaigns. In WTMC 2023.
[5] E. Chiapponi (2021). Scraping Airlines Bots: Insights Obtained Studying Honeypot Data. In International Journal of Cyber Forensics and Advanced Threat Investigations.

**Check them here:**

amadeus

amadeus

# Backup slides